化学指纹图谱的相似性测度及其评价方法

程翼宇* 陈闽军 吴永江
(浙江大学材料与化工学院 制药工程研究所 杭州 310027)

摘要 提出化学指纹图谱相似性测度概念,并以峰数弹性、峰比例同态性和峰面积同态性为评价指标,用于多角度评价化学指纹图谱相似性测度优劣,根据这些评价指标,用计算机仿真方法研究比较了 6 种相似性测度,结果表明夹角余弦测度用于度量指纹图谱间谱峰比例的波动较适宜,峰匹配测度可较灵敏地检测小峰个数波动,而欧氏距离测度则具有较好的综合评价能力. 最后,采用实测的参麦注射液色谱分析谱图,研究考察了上述计算机仿真实验结果,证明仿真结果符合实际情况,这表明本研究方法可用于评价化学指纹图谱相似性测度.

关键词 化学指纹图谱,指纹图谱相似性测度,中药质量控制

Measures for Determining the Similarity of Chemical Fingerprint and a Method of Evaluating the Measures

CHENG, Yi-Yu* CHEN, Min-Jun WU, Yong-Jiang
(*Institute of Pharmaceutical Engineering, College of Materials Science & Chemical Engineering, Zhejiang University, Hangzhou 310027*)

Abstract For determining similarity of chemical fingerprint, similarity measures were presented. A method including three indexes (*i. e.* elasticity of peak number, homostasis of proportion among peaks and homostasis of peak area) was proposed to evaluate the performance of the measures on different aspects. According to the three proposed indexes, six different measures were evaluated with chemical fingerprints simulated by computer. The results indicated that cosine of the angle is much suitable for assessing the fluctuation of proportion between the peaks. The matching degree of peaks is very sensitive in detecting the variation of number of small peaks. The Euclidean distance has preferable integral measurement capability. The actual chemical fingerprints of Shenmai injection were used to check the correctness of the above conclusion from simulation experiments, and the results showed it accords with reality. It could be concluded that the proposed method could be used for selecting the measures for determining the similarity of chemical fingerprint.

Keywords chemical fingerprint, similarity measure of chemical fingerprint, quality control of Traditional Chinese Medicine

化学指纹图谱分析是一种从整体上研究复杂物质体系的技术工具,已经在环境保护[1]、石油勘探[2]、食品评价[3]等许多领域中得到广泛应用. 由于它具有指纹特征分析、谱图整体分析、宏观推断分析等特点,适合于分析复杂化学物质组成的稳定性,故成为当今中药乃至天然产物质量评价研究领域的前沿课题,引起分析化学界的高度关注.

中药往往是由众多化学组分所组成的复杂物质体系,药材的生物多样性和制药过程参数的波动均

* E-mail: chengyy@zju.edu.cn
Received May 21, 2002; revised July 24, 2002; accepted August 17, 2002.
国家重点基础研究发展规划(No. G1999054405)及国家"十五"重大科技攻关计划(No. 2001BA701A01)资助项目.

会引起药品化学组成的变化,传统的测定一个或几个化学成分含量的方法无法对中药质量作出整体分析,从而形成药物分析学科的一大难题.近来人们转向化学指纹图谱分析技术,试图发展一种用于综合评价天然产物质量的新方法.化学指纹图谱分析不等同于一般的仪器分析,它通常由指纹图谱获取及指纹图谱鉴别计算两部分所组成,前者是指采用色谱、光谱以及联用等仪器分析方法,获取能表征样品化学组成特征的组分群体分析谱图或图像;后者是运用计算分析技术对所获得的谱图数据进行处理,通过筛选和简化,获得专属、宏观、整体的化学特征综合信息,并对样品化学组成的总体波动情况进行估测.因此,在获取了样品指纹图谱后,如何计算并评定各批次产品指纹图谱间的整体相似程度,定量描述它们各自化学组成的差异性和波动情况极为重要,即指纹图谱相似性计算是化学指纹图谱分析中的核心环节之一.为此,迫切需要研究化学指纹图谱相似性测度并建立评价相似性测度的方法.

本文研究化学指纹图谱相似性测度及其评价方法,提出以峰数弹性(elasticity of peak number)、峰比例同态性(homostasis of proportion among peaks)及峰面积同态性(homostasis of peak area)三个指标评价指纹图谱相似性测度的优劣.根据上述评价指标,用计算机仿真方法考察和比较了6种相似性测度,并以参麦注射液为例,用实测指纹图谱考核了仿真实验所得结论,证明本文方法可以用于化学指纹图谱相似性测度的评价.

1 化学指纹图谱相似性计算理论

在进行化学指纹图谱相似性计算前,先要建立能表征中药产品化学组成特征的标准指纹图谱,才能通过计算待测样品指纹图谱与标准指纹图谱间的相似度,对产品质量波动情况作出整体分析和评价.一般而言,通过比较计算各批次样品指纹图谱中的谱峰与标准指纹图谱中对应谱峰间的相似度,可度量出它们化学组成的相似程度.

1.1 化学指纹图谱的相似性计算

化学指纹图谱相似性计算就是通过量化比较各张指纹图谱上互相对应谱峰的异同,计算出各指纹图谱间的相似程度.它包括两个步骤:确定用于量化比较的谱峰测量参数(如峰面积、峰高或电平等);计算指纹图谱间相似度.指纹图谱相似度可以采用一个或几个指纹峰进行比较计算,也可以采用指纹图谱上所有谱峰进行比较计算,为较全面度量中药产品化学组成的整体波动情况,一般倾向采用后一种计算方法.

为量化比较谱峰测量参数值,从而量测出指纹图谱相似度,必须建立科学规范的化学指纹图谱相似性测度.

1.2 化学指纹图谱相似性测度

化学指纹图谱相似性测度是指用于表征指纹图谱间相似性的度量衡,即量测指纹图谱相似度的计算比较式.

这里,定义指纹图谱中的各指纹峰构成一个模式空间,即视各指纹峰测量参数值为某模式向量的一组元素;令指纹峰数为模式空间的维数,将一张指纹图谱在模式空间中表达为一个模式向量,从而将 N 张指纹图谱间相似性的量化比较转化为计算模式空间中 N 个模式向量的相似度,由此可将度量模式向量相似性的测度用作化学指纹图谱相似性测度.

设某产品指纹图谱由两个谱峰组成,以这两个谱峰的峰面积值 x_1, x_2 构成模式向量空间坐标,用向量 $X_s = [x_{s1}, x_{s2}]$ 表征标准指纹图谱,而用向量 $X_t = [x_{t1}, x_{t2}]$ 表征待测指纹图谱,如图1所示.这样,待测指纹图谱与标准指纹图谱间的相似度就可通过计算向量 X_s 与 X_t 间的相似度来获取.

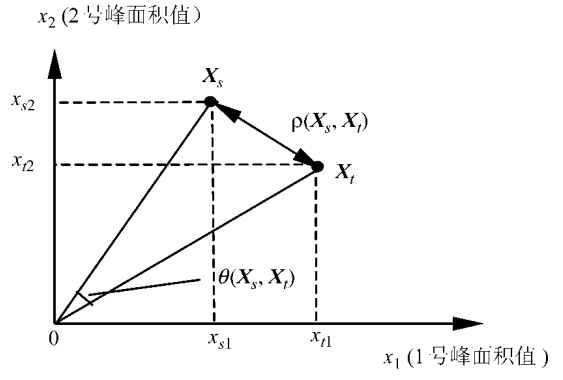


图1 用二维模式向量表征具有两个谱峰的指纹图谱

Figure 1 Representation of a fingerprint with two peaks using a vector of two dimension

模式向量 X_s 与 X_t 间相似度可以用向量间距离或夹角计算得到.根据距离 $\rho(X_s, X_t)$ 的非增函数设立相似性测度,如下式:

$$S(X_s, X_t) = e^{-\rho(X_s, X_t)/\rho(X_s, 0)} \quad (1)$$

其中 **0** 为零向量,$\rho(X_s, X_t)$ 为向量 X_s 与 X_t 间的空间距离,$\rho(X_s, 0)$ 为向量 X_s 与 **0** 间的空间距离,此

处空间距离的计算式可采用欧氏距离、绝对距离或者 Minkowsky 距离(本文取 $\lambda = 3$)等计算式,具体公式参见文献[4].

根据夹角 $\theta(X_s, X_t)$ 的余弦函数设立相似性测度,如下式:

$$S(X_s, X_t) = \frac{\sum_{i=1}^{2} x_{si} \cdot x_{ti}}{\sum_{i=1}^{2} x_{si}^2 \cdot \sum_{i=1}^{2} x_{ti}^2} \tag{2}$$

根据向量 X_s 与 X_t 的相关系数设立相似性测度,计算式为:

$$S(X_s, X_t) = \frac{\sum_{i=1}^{2} (x_{si} - \bar{x}_s) \cdot (x_{ti} - \bar{x}_t)}{\sqrt{\sum_{i=1}^{2} (x_{si} - \bar{x}_s)^2 \cdot \sum_{i=1}^{2} (x_{si} - \bar{x}_t)^2}} \tag{3}$$

其中 $\bar{x}_s$ 和 $\bar{x}_t$ 分别为模式向量 X_s 和 X_t 的平均值.

另外,还可根据指纹图谱间对应谱峰的匹配情况设立相似性测度,即用待测指纹图谱与标准指纹图谱中的共有谱峰数除以标准指纹图谱中谱峰总数,得到峰匹配度.

如果指纹图谱由三个谱峰组成,可在三维模式空间中用模式向量[x_1, x_2, x_3]表征指纹图谱;依此类推,就可在 N 维模式空间中用模式向量[x_1, x_2, …, x_N]表征由 N 个谱峰组成的指纹图谱. 类似地,将上述用于二维模式向量的各种相似性测度推广至 N 维后,含有任意谱峰数目的指纹图谱均可依此进行相似性计算.

1.3　化学指纹图谱相似性测度的评价方法

各种相似性测度的量测性能是不同的,故需建立一种评价方法来比较研究化学指纹图谱相似性测度的优劣.

为准确度量化学指纹图谱间相似性,所使用的相似性测度应该能够有效地量测出指纹图谱间的异同. 作者认为,两指纹图谱间的总体异同比较,应该包括指纹图谱间对应谱峰个数的差异、谱峰比例的差异以及总峰面积差异等三个方面. 为评价各种相似性测度对上述方面差异的量测能力,本文建立了如下三个指标:

(1) 峰数弹性:峰数弹性是用于评价某相似性测度检测指纹图谱间对应谱峰个数波动能力的指标,它可定义为相似性测度值变化的相对量与指纹图谱间谱峰个数差异的相对量之比,其计算式如下:

$$E = (\Delta S / S) / (\Delta N / N) \tag{4}$$

其中 ΔS, ΔN 分别为因待测指纹图谱中谱峰增减所产生的相似性测度值变化量和谱峰增减数量,S 为没有谱峰增减时两指纹图谱间的相似度,N 为待测指纹图谱中谱峰数. 假定与标准指纹图谱相比较,待测指纹图谱中有一谱峰增减,采用某相似性测度代入式(4)就可计算得到一个 E 值;如果待测指纹图谱有 N 个谱峰,每个谱峰依次增减一次就可计算得到 N 个 E 值,可取其中最小者作为该相似性测度的峰数弹性值. 峰数弹性值越大,表明该测度检测指纹图谱间对应谱峰数差异的能力越强. 在各种相似性测度中,小峰对相似度的贡献往往相对较小,峰数弹性的取值总是取决于小峰增减所得到的计算值,故峰数弹性也可视作某相似性测度检测小峰增减能力的评价指标.

(2) 峰比例同态性:同态性通常是指经映射后的数值大小揭示映射前样本之间所存在某种关系的能力. 此处,峰比例同态性是一个反映相似性测度对指纹图谱间谱峰比例关系变动情况检测能力的评价指标. 若用模式向量 $X = [x_1, x_2, \cdots, x_N]$ 描述某指纹图谱,其中 x_i 为第 i 个谱峰的面积值,N 为谱峰数;用 $Y = X / \sqrt{\sum_{i=1}^{N} x_i^2}$ 表示指纹图谱上谱峰比例,待测指纹图谱与标准指纹图谱间谱峰比例关系的差异量 ΔP 就可用下式计算:

$$\Delta P = \sum_{i=1}^{N} | y_{si} - y_{ti} | \tag{5}$$

式中下标符号与前述同. 峰比例同态性可以采用 ΔP 与相似性测度值变化量 ΔS 的相关系数进行计算,如下式:

$$H = f(\Delta P, \Delta S) \tag{6}$$

其中函数 $f(\cdot)$ 的计算式类似公式(3). 当 $H = 1$ 时,表示该相似性测度具有完全同态性,能够很好反映一张指纹图谱上各谱峰比例关系与另一张指纹图谱上各峰比例关系间的差异;若 $H = 0$,则表示该相似性测度完全没有同态性,不能够恰当反映两张指纹图谱间谱峰比例关系的差异.

(3) 峰面积同态性:峰面积同态性是用来反映

相似性测度对指纹图谱中所有谱峰的总面积波动情况检测能力的评价指标. 为保证该指标的独立性,应在指纹图谱的谱峰数和谱峰比例不变前提下,通过计算总峰面积变化量与相似性测度值变化量的相关系数得到峰面积同态性值,计算公式类似式(6),其中 ΔP 用总峰面积变化量代替,不再赘述.

2 结果与讨论

2.1 数值仿真研究

采用计算机仿真方法产生一系列反映谱峰变化情况的指纹图谱,用于评价 6 种相似性测度的量测性能. 考虑到目前常用指标性成分含量或大类成分总含量来检测中药产品质量,作为对照比较,也将这两种测度(分别简称为指标成分相似度与大类成分相似度)一起进行考察.

计算机仿真指纹图谱采用高斯模型产生,数学模型表达式为:

$$h(t) = \sum_{i=1}^{N} \frac{A_i}{\sqrt{2\pi}\,\sigma} \exp\left(- \frac{(t - t_i)^2}{2\sigma^2} \right) \qquad (7)$$

其中 N 为总峰数,t_i 为第 i 峰保留时间,A_i 为第 i 峰面积值,σ 为高斯函数标准偏差. 设 N 等于 8, t_i (i = 1, 2, …, 8) 分别为 2, 4, 6, 8, 10, 12, 14, 16, A_i (i = 1, 2, …, 8) 分别为 1, 1, 3, 5, 10, 20, 30, 30, σ 为 0.2,将这些参数代入式(7)可模拟产生 8 个谱峰. 用这 8 个谱峰组成仿真的标准化学指纹图谱,设第 8 个峰为指标成分峰. 通过人工扰动各谱峰面积产生大量仿真的待测化学指纹图谱,分别模拟总峰面积波动、峰比例波动、峰个数波动等各种谱峰变化情况.

在计算机仿真实验中,使用本文方法对各相似性测度进行评价,结果如表 1 所示. 由表 1 知,欧氏距离、绝对距离以及 Minkowsky 距离等三种距离型测度的评价结果没有显著差别,表明它们对谱峰各种波动情况的量测能力相近,故可选择最常用的欧氏距离为代表来讨论距离型测度;夹角余弦与相关系数两种测度的评价结果也基本接近,且夹角余弦在峰比例同态性指标上略优,故选其为代表参与相似性测度的比较研究.

由峰数弹性指标项的评价结果可知,对指纹图谱间小峰差异检测能力最强的测度是峰匹配度,其余依次为欧氏距离 > 大类成分相似度 > 夹角余弦 > 指标成分相似度. 这里,指标成分相似度的测试值为零,表明其不能检测化学指纹图谱中小峰个数的波动,这与实际经验相一致,因为指标成分相似度并不涉及指标成分峰之外的其它谱峰,这些谱峰的增减不会引起指标成分相似度的变化;夹角余弦测度的测试值也很低,提示它检测各指纹图谱间小峰差异的能力较差.

由峰比例同态性指标项的评价结果可知,对各指纹图谱间谱峰比例关系差异检测能力最强的测度是夹角余弦,其余依次为欧氏距离 > 大类成分相似度 > 指标成分相似度 > 峰匹配度. 这里,峰匹配度测试值为零,表明其不能检测各指纹图谱间谱峰比例的差异,这也与实际经验相吻合;大类成分相似度和指标成分相似度的测试值亦较低,表明两者对不同样品间化学组分比例差异的检测能力较差,这就解释了测定指标性成分含量或大类成分总含量的质控方法为何难以保证中药产品质量的稳定性.

由峰面积同态性评价结果可知,欧氏距离、大类成分相似度以及指标成分相似度等测度的测试值都接近或等于 1,表明它们对各指纹图谱间谱峰总面积波动情况的检测能力均较强;而夹角余弦和峰匹配度的测试值都为零,表明在谱峰数、谱峰比例关系不变条件下,这两种测度均无法检测出各指纹图谱间总峰面积的波动.

此外,由表 1 知,距离型测度在三个评价指标上的测试值都较大,表明它对各张指纹图谱间各类差异的检测能力均较强,具有较好的综合评价能力.

表 1 各相似性测度的评价结果

Table 1 The evaluation results of different similarity measures

评价指标	相似性测度							
	欧氏距离	绝对距离	Minkowsky 距离	夹角余弦	相关系数	峰匹配度	大类成分相似度	指标成分相似度
峰数弹性	0.168	0.080	0.168	0.016	0.032	1.000	0.080	0.000
峰比例同态性	0.798	0.653	0.803	1.000	0.973	0.000	0.535	0.455
峰面积同态性	0.991	0.991	0.994	0.000	0.000	0.000	1.000	1.000

2.2 实例研究

为验证上述计算机仿真实验所得结论,选择参麦注射液化学指纹图谱作为实例进行考察. 由某公司提供 2 个批次合格品及 1 个批次不合格品,用高效液相色谱测定它们的化学指纹图谱,如图 2 所示. 取其中 1 个批次合格品的谱图作为标准指纹图谱,用欧氏距离测度、夹角余弦测度以及峰匹配测度计算不同批次产品指纹图谱的相似度,以分析评价参麦注射液产品批次间质量稳定性,结果见表 2.

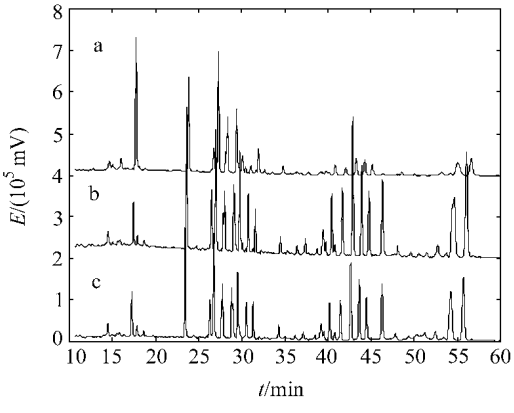


图 2 3 个批次参麦注射液化学指纹图谱

Figure 2 The chemical fingerprints of Shenmai injection obtained from three batches

a —Abnormal product; b,c —normal product

表 2 3 个批次参麦注射液产品指纹图谱相似性的计算结果

Table 2 The similarity of chemical fingerprints of Shenmai injection obtained from three batches

	夹角余弦	欧氏距离	峰匹配度
合格品	0.980	0.806	1.000
不合格品	0.745	0.434	0.760

由表 2 知,由于各种相似性测度对指纹图谱中谱峰各类变动情况的检测能力不相同,其相似性计算结果有较大差别. 如对合格品,夹角余弦测度的计算值为 0.980,表明其指纹图谱的谱峰比例与标准指纹图谱的谱峰比例相比基本相近,正如图 2 所示,2 个批次合格品指纹图谱中谱峰比例基本一致,表明计算结果符合实际. 欧氏距离测度的计算值为 0.806,这可能是由于指纹图谱中谱峰数、谱峰比例以及总峰面积的差异都会影响欧氏距离测度的计算,导致相似度偏低. 峰匹配测度的计算值为 1.000,提示其不能检测出各批次产品指纹图谱间谱峰比例关系的细微差异. 不合格品的夹角余弦和欧氏距离测度计算值与合格品的计算值存在明显差异,表明这两种相似性测度均能分辨合格品与不合格品. 不合格品的峰匹配测度计算值明显低于合格品的计算值,从图 2 可见,与标准指纹图谱相比,不合格品指纹图谱中确实有一些小峰缺失,提示该测度检测小峰增减的能力较强. 相比较而言,这三种测度中的夹角余弦测度更适于分析评价不同批次药品质量的稳定性. 由以上实例考察结果可知,计算机仿真实验所得结论符合药品分析的实际情况.

3 结论

根据峰数弹性、峰比例同态性和峰面积同态性等 3 个指标组成的相似性测度评价方法,考察了欧氏距离、绝对距离、Minkowsky 距离、夹角余弦、相关系数以及峰匹配度等 6 种相似性测度,计算机仿真实验结果表明三种距离测度性能相近,夹角余弦测度与相关系数测度的评价结果也接近;相比较而言,夹角余弦测度对于检测指纹图谱间谱峰比例关系的变动较为适宜,峰匹配测度能较灵敏地检测小峰个数波动,欧氏距离测度则具有较好的综合评价能力. 用 3 种相似性测度计算不同批次参麦注射液产品指纹图谱相似性的结果表明,夹角余弦测度更适于分析评价不同批次间药品质量的稳定性. 该实例分析结果证明,计算机仿真实验结论符合实际情况. 因此,本文方法可用于化学指纹图谱相似性测度的选择,有助于化学指纹图谱分析技术在中药质量控制等领域的推广应用.

References

1 Xu, S.; Zhang, Q.-H.; Wu, W.-Z; Li, W. *Chin. Sci. Bull.* **2000**, *45*(6), 578 (in Chinese).
(徐盛, 张庆华, 吴文忠, 黎雯, 科学通报, **2000**, *45*(6), 578.)

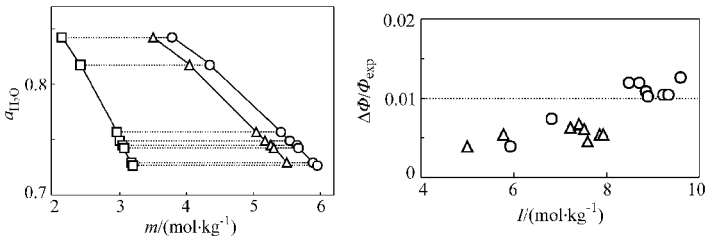
2 Devon, R.; Karlis, M. *Org. Geochem.* **1999**, *30*, 861.

3 Berente, B.; Garcya, D. C. *J. Chromatogr., A* **2000**, *871*, 95.

4 Shen, Q.; Tang, L. *Introduction to Pattern Recognition*, Science and Technology of National Defense University Press, Changsha, **1991**, p. 29 (in Chinese).
(沈清, 汤霖, 模式识别导论, 国防科技大学出版社, 长沙, **1991**, p. 29.)

(A0205218 SHEN, H.; DONG, H. Z.)

Isopiestic Studies of Synthetic Salt Lake Brine System Li-Na-K-Mg-Cl-SO_4-H_2O at 25 °C and Applications of Ion-Interaction Model

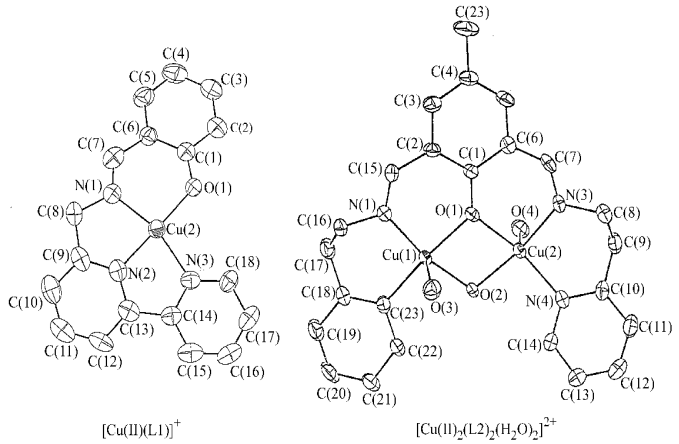


The water activities and osmotic coefficients were determined for the synthetic interstitial brine systems of Yi Li-Ping and Dong Tai. salt lakes from diluted to saturated concentrations at 25 °C. The thermodynamic properties were compared between the two brines, a cause of formation of the differences in concentrations was indicated. The calculated osmotic coefficients and the saturation indices [ln (k/k)] using the extended ion-interaction model agree with the experimental data in this work and with the results from the field observation and the evaporation-crystallization experiment reported.

YAO, Yan; SONG, Peng-Sheng; WANG, Rui-Ling; LONG, Guang-Ming

Acta Chimica Sinica **2002**, *60*(11), 2004

Crystal Structures and Spectroscopic and Electrochemical Properties of Schiff Base Cu(II) Complexes

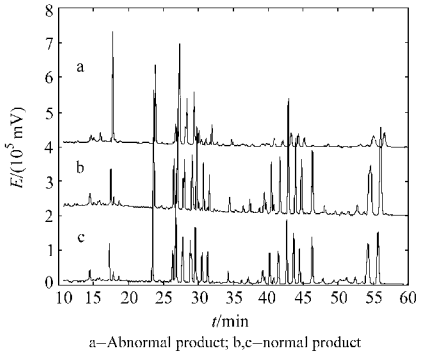


Two Schiff base Cu(II) complexes, [Cu(II)(L1)]ClO_4 (**1**), [Cu(II)$_2$(L2)-$(H_2O)_2$]$(BF_4)_2$ (**2**), were prepared and their structure were determined by X-ray diffraction.

YIN, Ye-Gao; LI, Dan; WU, Tao

Acta Chimica Sinica **2002**, *60*(11), 2011

Measures for Determining the Similarity of Chemical Fingerprint and a Method of Evaluating the Measures



The measures for determining similarity of chemical fingerprint were presented and a method including three indexes was proposed to evaluate their performance. From the results of simulation experiment and the investigation of the actual chemical fingerprint, it could be concluded that the proposed method could be used for selecting the similarity measures.

CHENG, Yi-Yu; CHEN, Min-Jun; WU, Yong-Jiang

Acta Chimica Sinica **2002**, *60*(11), 2017